PMut2: a web-based tool for predicting pathological mutations on proteins

V. López-Ferrando[1], X. de la Cruz[2], M. Orozco[1, 3], J.L. Gelpí[1, 3]
1. Joint IRB-BSC-CRG programme for Computational Biology. Life Sciences Dept. Barcelona Supercomputing Center.
2. Vall d'Hebron Research Institute.
3. Dept. Biochemistry and Molecular Biology. Universitat de Barcelona.

Abstract- *Amino acid substitutions in proteins can result in an altered phenotype which might lead to a disease. PMut2 is a method that can predict whether a mutation has a pathological effect on the protein function. It uses current machine learning algorithms based on protein sequence derived information. The accuracy of PMut2 is as high as 82%, with a Matthews correlation coefficient of 0,62. PMut2 predictions can be obtained through a modern website which also allows to apply the same machine learning methodology that is used to train PMut2 to custom training sets, allowing users to build their own tailor-made predictors.*

I. INTRODUCTION

Assessing the impact of amino acid mutations in human health is an important challenge in biomedical research. As sequencing technologies are more available, and more individual genomes become accessible, the number of identified variants has dramatically increased. PMut, released back in 2005 [1], has been one of the popular predictors in this field. PMut was a neural-network-based classifier using sequence data to provide a pathology score for point mutations in proteins.

PMut2 is a new, revised, and much more powerful version of the predictor. It introduces the use of state-of-the-art machine learning algorithms and an updated training set based on SwissVar [2]. It achieves an accuracy of 82% and a Matthews correlation coefficient (MCC) of 0.62. PMut2 includes a fully featured training and validation engine that can be optimized to generate predictors adapted to user specific training sets. The engine is implemented in Python using MongoDB engine for data management. It has been adapted to run at the HPC level to cover large scale annotation projects.

II. METHODS

common machine learning methods. First of all, a training set of mutations annotated as either neutral or pathological must be established. For each of these variants a set of numerical features are computed to best describe them. Finally, a model is selected and trained using the training set and the computed features.

Training set

PMut2 is trained using the manually curated variation database SwissVar (as of December 2015), which contains ~28,000 disease and ~38,000 neutral mutations on ~12,500 proteins.

Features computation

Over 150 numerical features are computed for each mutation. They account for 1) physical property differences between wild type and mutated amino acids, 2) protein interactome information and 3) amino acid conservation. The conservation features are derived from local searches over UniRef100 and UniRef90 clusters [3] using PSI-BLAST [4] and multiple sequence alignments using Kalign2 [5].

Model selection

In order to choose a model for the predictor, different classifiers were tested with different parameter configurations. Random Forest is chosen as it presents the best predictive power and good computing speed.

From the 150 features computed, only 12 of them were selected via an iterative algorithm to be part of the final predictor, as seen in Figure .

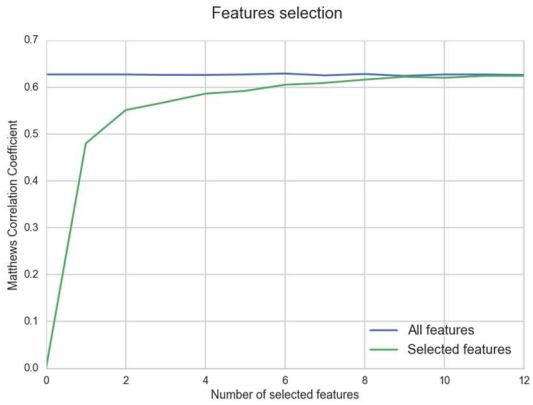


Figure 1. Iterative features selection. With 12 features the model has the same precision as with all 150 features.

III. RESULTS

using 10-fold cross validation with 50% protein identity exclusion in the different folds. This method allows us to estimate the predictor performance when faced with a protein different to the ones in its training set. The predictor obtained a Matthews correlation coefficient of 0,620 ± 0,02 and an Area under the ROC curve of 0,808 ± 0,01.

Predictor comparison

The predictor performance is compared to other popular predictors in the field using new mutations that were added to SwissVar after the model

training (January – March 2016). Table 1 holds the results of this comparison.

The meta-predictor PROVEAN presents the best performance, with an MCC of 0,479; PMut2 follows with an MCC of 0,469. This comparison also outlined how Classic PMut performs poorly.

Table 1. Comparison of different predictors by predicting 573 mutations added to SwissVar in January – March 2016.

Predictor	Coverage	Accuracy	Sensitivity	Specificity	AUC	MCC
PROVEAN	98,3	0,762	0,814	0,819	0,740	**0,479**
SIFT	97,6	0,835	0,835	0,774	0,689	0,395
Polyphen	98,6	0,763	0,884	0,777	0,714	0,463
Condel	95,3	0,758	0,843	0,794	0,724	0,462
Classic PMut	52,9	0,551	0,507	0,791	0,585	**0,154**
PMut2	100	0,743	0,746	0,839	0,742	**0,469**

Web application

A web application allows use of PMut2 predictor and provides other useful functionality (Figure 2). A comprehensive repository of precomputed predictions yields instant access to all possible mutations in all human proteins in UniProt (a total of 803,743,460 variants on 109,106 proteins). Fig. 4 shows one of such proteins in the repository, Fig. 5 is the results page of an analysis of a list of mutations and Fig. 6 is the page of a custom predictor that has been trained using a given training set.

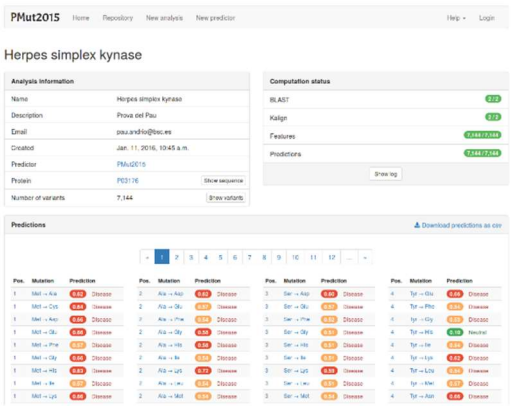
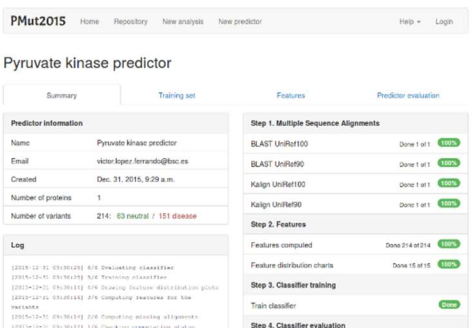


Figure 2: Web application Home page summarizing the three use cases:1) Browser or search the repository, 2) Get predictions of given mutations and 3) Train a tailor-made predictor.

are predicted and their predictions can be observed in the 3D structure

Figure 3: Example of protein in the Repository. All possible mutations



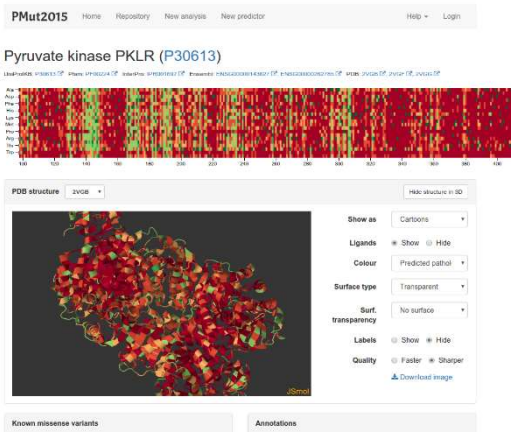


Figure 4: Example of analysis result. After all computations complete,each variant has a corresponding predicted pathology score.

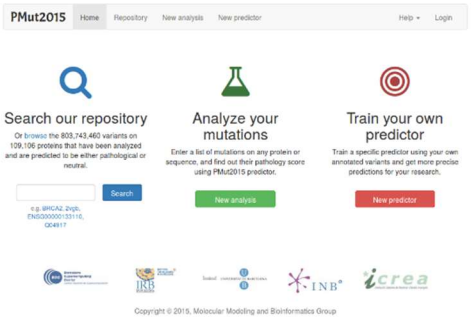


Figure 5: Example of custom predictor for protein Pyruvate Kinase. A training set consisting of a list of variants annotated as disease or neutral was submitted.

IV. CONCLUSION

Using state-of-the-art machine learning algorithms we were able to train a predictor that matches the predictive power of the most popular predictors in the field. Furthermore, the methods used have been automated and offered to the research community via a user-friendly website.

REFERENCES

[1] Ferrer-Costa, et al. «PMUT: A Web-Based Tool for the Annotation of Pathological Mutations on Proteins». Bioinformatics 21, num. 14 (15 July 2005): 3176-78.

[2] Mottaz, et al. «Easy Retrieval of Single Amino-Acid Polymorphisms and Phenotype Information Using SwissVar». Bioinformatics 26, num. 6 (15 March 2010): 851-52.

[3] Suzek, et al. «UniRef: Comprehensive and Non-Redundant UniProt Reference Clusters». Bioinformatics 23, num. 10 (15 May 2007): 1282-88.

[4] Altschul, et al. «Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs». Nucleic Acids Research 25, num. 17 (9 January 1997): 3389-3402.

[5] Timo Lassmann, Oliver Frings, and Erik L. L. Sonnhammer, "Kalign2: High-Performance Multiple Alignment of Protein and Nucleotide Sequences Allowing External Features," Nucleic Acids Research 37, no. 3 (January 2, 2009): 858–65, doi:10.1093/nar/gkn100